DOCUMENT RESUME

ED 148 876 TM 006 766

AUTHOR Modu, Christopher C.; Stern, June
TITLE The Stability of the SAT-Verbal Score Scale.
INSTITUTION College Entrance Examination Board, New York, N.Y.
SPONS AGENCY Educational Testing Service, Princeton, N.J.
PUB DATE May 77
NOTE 36p.; For related documents, see TM 006 750-776 and TM 006 805; For complete report, see TM 006 516
AVAILABLE FROM College Board Publication Orders, Box 2815, Princeton, New Jersey 08541 (Item No. 2517213, $3.00)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS College Bound Students; *College Entrance Examinations; Educational Problems; *Educational Trends; *Equated Scores; Performance Factors; Scores; Secondary Education; Statistical Analysis; Tables (Data); Testing Problems; *Test Reliability; Trend Analysis; Verbal Ability; *Verbal Tests
IDENTIFIERS *Scholastic Aptitude Test; *Test Score Decline

ABSTRACT

To assess the stability of the Scholastic Aptitude Test verbal score scale SAT--V, 1963 and 1973 forms of the SAT--V were administered in counterbalanced order to spaced samples of the same group. The 1973 scores were placed on the reporting scale used for the 1963 form. The experimentally derived scores on the 1963 scale were compared with their corresponding scores on the 1973 scale for candidates of the same ability levels, in order to estimate the degree of scale shift. These findings confirmed an upward scale drift which gave the 1973 candidate group an average of 8 to 10 points higher than they would have earned had these experimental equating results rather than the operational equating results been used in reporting the 1973 scores. The scale drift was not uniform over the extent of the scale; it was found to increase as scores decreased from 600 to 200. Little scale drift was noticed over the 10-year span for scores of 650 and above. It would appear that the reported declines in mean SAT--V scores from 1963 to 1973 are about 8 to 10 points smaller than they would have been had the scale been completely stable. (Author/EVH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED148876

The Stability of the SAT-Verbal Score Scale

Christopher C. Modu
June Stern
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
College Entrance Examination Board
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 766

Prepared for the Advisory Panel on
the Scholastic Aptitude Test Score Decline,
jointly sponsored by the
College Board and Educational Testing Service

May 1977

This appendix is one of 27 that have been published individually to supplement On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline. The report itself is available at $4 from College Board Publication Orders, Box 2815, Princeton, New Jersey 08540. Additional copies of this appendix, at $3 per copy, as well as a price list for all the appendixes, can be obtained from the same address. Payment must accompany any order not submitted on an institutional purchase order.

Copyright © 1977 by College Entrance Examination Board. All rights reserved.

Printed in the United States of America.

ACKNOWLEDGMENTS

Grateful acknowledgments are due Bruce K. Eckland, William B. Schrader and Ledyard R Tucker for their invaluable suggestions for improving the design of the study, to Rebecca Bullock for her significant contribution to the completion of the analysis phase of the study, and to William H. Angoff, George H. Hanford, Rex Jackson, Frederick M. Lord, Gary L. Marco, and E. Elizabeth Stewart for their careful reviews of the initial draft of this report.

This study was funded by the Joint Staff Research and Development Committee of the College Entrance Examination Board and Educational Testing Service.

ABSTRACT

This study was designed to determine the extent of drift in the Scholastic Aptitude Test (SAT) score scale between 1963 and 1973 for the SAT-verbal sections only. In a previous study, scores on two old forms were equated to a 1973 form and, through that form, to the College Board scale by means of sets of items common to the new and old forms. In the present study, pairs of 1963 and 1973 forms of the verbal sections were administered in counterbalanced order to spaced samples of the same group, each candidate taking a 1963 and a 1973 form. The obtained scores were used to place the 1973 scores on the reporting scale used for the 1963 form. The experimentally derived scores on the 1963 scale were then compared with their corresponding scores on the 1973 scale for candidates of the same ability levels in order to estimate the degree of scale shift.

The findings of the present study confirm an upward scale drift that gave the 1973 candidate group an average of 8 to 10 points higher than they would have earned had these experimental equating results been used in reporting the 1973 scores rather than the operational equating results. The scale drift observed for the present study was not uniform over the extent of the scale; it was found to increase as scores decreased from 600 to 200. Little scale drift was noticed over the 10-year span for scores of 650 and above. On the basis of the evidence from this study it would appear that operational equating underestimates the mean verbal score declines by about 8 to 10 points over the 10-year span covered by the study. If one of the two conversion lines for the May 1963 operational equating is ignored because of the inconsistency of its results, and if only the second conversion line is considered along with the results for the December 1963 equatings, the scale drift would be about 11 to 13 points rather than 8 to 10.

INTRODUCTION

Considerable interest continues to be expressed in the stability of the verbal and mathematical scales of the Scholastic Aptitude Test. Test-score recipients wish to be reassured that a given scaled score represents the same level of performance on different forms of the SAT. Since scores on each test are placed on scale through form-to-form equating, only by establishing the consistency of the scale system or determining the extent of drift in the scale over a period of time are users able to make meaningful comparisons of populations tested during those years.

The present study was designed to ascertain the extent of drift that may have occurred in the verbal score scale between 1963 and 1973. The study was limited to the verbal sections of the test partly because the mean SAT score decline during the 10-year period covered by the study was greater for verbal than for mathematical scores and partly because the content specifications of the 1976 operational mathematical test differ from those of the 1973 test and it was therefore felt that the candidates used for the study might have been ill-prepared to handle the 1963 or 1973 mathematical test.

In a recent study of the stability of the SAT score scale, Modu and Stern (1975) concluded that the scale had not remained invariant over the 10-year period, 1963-1973. The results of their study showed that the candidates who took the December 1973 form of the SAT earned higher scores than they would have obtained had they taken the December 1963 or 1966 form. The amount of the difference was estimated to be 14 points on the verbal test and 17 points on the mathematical test. Although Modu and Stern followed the same method used by Stewart (1966) in her study--which reached the conclusion that the verbal scale was relatively stable between 1953 and 1963--their report underscored the following methodological considerations that might have contributed to the observed discrepancy between the scores obtained by a candidate on the 1963 or 1966 form and the scores a 1973 candidate of equivalent ability would have earned had he taken those forms:

1. The effect of differences in the performance of the old and new form samples;
2. the differences resulting from the use of Tucker and Levine methods (Angoff, 1971) for operational equatings;
3. the effect of the methods of obtaining test reliability estimates used for the Levine equating procedure;
4. the effect of the choice of equating sections; and
5. the effect of the choice of the old form to which the new form is equated.

Thus, both the Stewart and the Modu and Stern studies are subject to the criticism that they used the same methodology employed in operational procedures that may have led to the reported scale shifts. Hence, the present study was designed to verify the SAT-verbal scale drift during the same 10-year period covered by Modu and Stern in their previous study by employing a different experimental procedure from the one used in earlier studies.

METHOD

Design of the Study

The procedure followed in this study involved the experimental equatings of the raw scores on the 1973 forms to those of the 1963 forms. The resulting raw-to-scaled score conversions were then compared with the scaled scores that were derived when the 1973 forms were first administered. If the score scale had remained stable, a given raw score on the 1973 test, for example, would have converted to the same scaled score on the 1963 or 1973 scale.

Unlike the method used by Stewart (1966) or Modu and Stern (1975), in which a recent form was administered along with the common equating sections from two earlier forms, the present study called for the simultaneous administrations of pairings of 1963 and 1973 verbal tests to four spaced samples of students in each participating school in the following order:

Sample A: May 1963 test (Y1); March 1973 test (X1)
Sample B: March 1973 test (X1); May 1963 test (Y1)
Sample C: December 1963 test (Y2); November 1973 test (X2)
Sample D: November 1973 test (X2); December 1963 test (Y2)

Thus, each student took one 1963 and one 1973 test, with each pair of 1963 and 1973 tests given in counterbalanced order to two different samples. The four test forms were, therefore, paired for printing in four test books as Y1-X1, X1-Y1, Y2-X2 and X2-Y2 and then administered in sequence (spiraled) in each school (i.e., as Y1-X1, X1-Y1, Y2-X2, X2-Y2, and repeat).

The experimental equating methods used in the present study for placing scores from the 1973 tests on the 1963 scale are described below. However, since the design of the study did not require the use of common equating sections, the possible biases listed under items (1) to (4) in the preceding section were eliminated. Also, the methodological weakness of the previous study indicated under item (5) was partially controlled by using two 1963 and two 1973 forms for the equatings instead of only one pair of 1963 and 1973 tests. Ideally, one should have averaged the results of equatings based on pairs of all four 1963 and five 1973 forms, had it been administratively possible to give all nine forms to random samples of the same group. Since the various SAT forms are built to parallel specifications--an assumption that will be verified in the present study--the variations in performance attributable to different pairings of 1963 and 1973 tests should be minimal when the tests are administered to random samples of the same group. Thus, the selection of two 1963 and two 1973 tests is designed (1) to replicate the experimental equating of a 1963 and a 1973 form used in the previous study, and (2) to determine the extent of scale drift based on an average of two independent pairings of test forms administered in the 10-year span.

SAMPLES

Sixty-six high schools in Connecticut, New Jersey, and Pennsylvania that test 100 or more candidates for the SAT each spring participated in the study. Virtually all the schools had mean scores that approximated that of the 1973 national SAT candidates. Motivation for taking the SAT in an experimental administration was partially secured by limiting the tests to college-bound juniors who were planning to take the SAT as juniors in June 1976 or as seniors in the fall of 1976 and might, therefore, benefit from the experience. Additional motivation was also assured by providing rosters of scaled scores obtained on the 1973 test to participating schools for release to students within six weeks of the test administration.

A total of 3,174 students participated in the study. Each of them took a 1963 and a 1973 test during a 2-1/2 hour testing period. Although student participation in the selected schools was voluntary, the schools were encouraged to ensure that the sample of about 40 to 60 students from each school represented the full range of abilities within the SAT-taking population for that school.

Between two and four cases were dropped at random from each of the four samples (A, B, C, and D) in order to reduce each sample to a maximum multiple of five cases as required by the item-analysis computer programs. These reduced samples were then used for placing the scores of 1973 verbal tests on the scale of the 1963 tests.

The raw-score summary statistics for the 1973 and 1963 SAT candidates compared to those for the study samples who took corresponding test forms are presented in Table 1. As can be observed from the table, the samples who took the 1973 tests first obtained higher mean scores than the other samples who took the same tests second. A statistically significant difference ($p < .01$) was observed between the means for the samples that took Test X1 first or second but not between those of the samples that took each of Tests X2, Y1, and Y2 first or second. The combined study samples obtained a higher mean score than the 1973 SAT candidates on Test X1, whereas the combined samples obtained a lower (but nonsignificantly different) mean score on Test X2 than the 1973 SAT candidates. In general, the study samples are of comparable or slightly higher ability than corresponding SAT candidate groups on the 1973 tests. By contrast, all four study samples obtained substantially lower mean scores (averaging 4 to 6 raw-score points) on the 1963 tests than the 1963 SAT candidates. These data emphasize the reality of mean score decline over the 10-year span covered by the study, since the study samples are comparable with the 1973 SAT candidates, as has already been pointed out.

Because there was some concern about the representativeness of the experimental samples compared with that of the operational SAT candidate groups, each sample was subdivided into high- and low-ability groups according to their self-reported class ranks. The equatings based on the total experimental groups for each pair of 1963 and 1973 tests were then repeated separately for high- and low-ability groups in order to assess the difference that any possible variation of the experimental sample from national SAT candidate groups might have had on the equating results.

Tests

As indicated earlier, two 1963 and two 1973 verbal tests were reprinted in four test books for the present study. Each test book contained a 1963 and a 1973 form, with each test divided into two separately timed sections of 30 and 45 minutes, thus resulting in an actual testing time of 150 minutes per candidate. There were no changes in either the timing or the number of items for the SAT-verbal during the 10-year period covered by the study nor were there any variations in format, content, and statistical specifications of the tests that could have seriously affected candidate performance. See Appendix I for the content specifications of the verbal tests.

Because the 1963 and 1973 tests were readministered to high school students in 1976, it was considered necessary to include in the design for the study item analysis procedures that might reveal, from the scatterplots of item-difficulty indices observed on the 1963 tests for (1) 1963 SAT candidates, and (2) the 1976 study samples, whether the 1963 test questions are as appropriate for the 1976 study samples. A similar analysis was conducted for the 1973 tests.

Table 1. Raw Scores Obtained by 1973 and 1963 SAT Candidates Compared with Those of Study Samples Who Took Corresponding Verbal Tests

Test form	SAT candidates	Sample taking test first	Sample taking test second	Combined samples
March 1973 (X1)				
N	120,593	790	790	1,580
Mean	33.1	35.7[a]	34.1	34.9[b]
S.D.	17.2	17.1	16.1	16.6
November 1973 (X2)				
N	377,846	795	785	1,580
Mean	36.0	36.0	34.9	35.4
S.D.	17.1	16.0	16.1	16.1[b]
May 1963 (Y1)				
N	250,905	790	790	1,580
Mean	38.2	34.1	34.5	34.3[b]
S.D.	16.8	16.9	17.6	17.2
December 1963 (Y2)				
N	484,705	785	795	1,580
Mean	41.3	35.1	34.9	35.0[b]
S.D.	17.1	16.8	16.9	16.8

a. Significantly different at $p < .01$ from corresponding statistics for the sample taking test second.
b. Significantly different at $p < .01$ from corresponding statistics for SAT candidates.

Equating Methods

The linear method described fully as Design II by Angoff (1971, p. 573) and as Case III by Lord (1950) was used for the experimental equating. This method assumes that:

1. The first of a pair of tests administered to a sample of examinees may have a practice or fatigue effect on the results obtained for the test taken second. This assumption requires that the pair of tests be administered in counterbalanced order to random halves of the examinees. Adjustments are made for the estimated practice or fatigue effect as part of the equating procedure.
2. The shape of the distribution of raw scores for one test does not differ markedly from the shape of the distribution of scores for the second test.
3. The pair of tests being equated are parallel forms.

After making adjustments for the estimated practice or fatigue effect, the means and standard deviations for the scores earned on each of the two forms were set equal to each other, and the slope and intercept of the line relating the scores on one form (1973) to the scores on the other (1963) were then derived.

In addition to the above linear procedure, a curvilinear method of equating, the equipercentile method (Angoff, 1971, p. 581), was used to equate scores on the pair of tests taken by each of the four samples, A, B, C, and D. The equatings were then averaged for Samples A and B combined in order to place the March 1973 test on the reporting scale for May 1963, and similarly for Samples C and D combined for the November 1973 test. The equipercentile equating results on the 1963 scale for the average of the two 1973 tests provided an additional check on the linear equating results.

Score Equating for High- and Low-Ability Groups

As an added check on our equating results, the linear equating method was applied to subsamples of differing ability levels. By means of a self-reported class rank indicated by the student at the time of the experimental administration (see Appendix VI), the samples were divided into two groups: (1) students who indicated that they stood in the top fifth of their high school class, and (2) students who responded that they were in one of the four categories below the top fifth. Results from these equatings were compared with those obtained previously for all students in each of the four samples.

RESULTS AND DISCUSSION

Appropriateness of the 1963 and 1973 Verbal Items for 1976 Study Samples

The equating results of this study could be affected by differential changes in the difficulty level of some of the items for the 1963 or

1973 SAT candidates versus the 1976 experimental samples. If, for example, some of the items on the 1963 test had become obsolete for the 1976 experimental samples and had changed radically in difficulty from the results obtained when the test was originally taken by the 1963 candidate group, an apparent shift in the score scale could have resulted from the differential changes in item difficulties that could have altered the mean scores for the experimental equating samples. Analyses of the item-difficulty indices were, therefore, carried out to determine whether any items in the 1963 and 1973 tests were inappropriate for the 1976 study samples.

As can be noticed from the scatterplots of the item-difficulty indices (deltas) for the 370-case samples of the 1963 candidate groups versus the 1976 experimental samples (see Appendixes II and III), the delta plots form elongated ellipses with plot correlations of .964 for both the May 1963 and the December 1963 test forms. Plots for the 1070- and the 1995-case samples of the March and November 1973 test forms, respectively, versus the 1976 experimental samples (see Appendixes IV and V) form tighter ellipses than those of Appendixes II and III. Their plot correlations of .989 and .994, which are higher than those of the 1963 tests, may be due to the fact that the 1976 samples are more like the 1973 than the 1963 candidate group.

Items that fall on the major axis of the ellipse are of relatively equal difficulty for both groups; points above the line represent items that are relatively more difficult for the 1963 sample, whereas those below the line represent those which are relatively more difficult for the 1976 study samples. No item type appears to be relatively easier or more difficult for any one group, since all item types straddle both sides of the major axis of the ellipse. One can observe from the displacement of the ellipse of points to the right of the equal-difficulty line for both groups (i.e., the broken line that forms a 45-degree angle with the vertical axis) that the 1976 study samples are not as able as the 1963 candidate groups. The relative difficulty of the 1963 tests for the SAT candidate groups and the study samples is further confirmed by the observed difficulty indices, delta, of 12.6 and 12.4 for the May 1963 and December 1963 groups, respectively, compared to 13.1 and 13.0 for the two study samples, A and C, who took these tests first. In comparison with the 1973 candidate groups, the 1976 study samples are seen from the scatterplot in Appendix IV to be slightly more able than the March 1973 candidate group, whereas the study samples are of comparable ability with the November 1973 candidate group in view of the virtual collinearity of the major axis of the delta plots with the 45-degree, equal-difficulty line.

A couple of items in the plots for 1963 tests that might be considered more difficult for one sample relative to the rest of the items in the test were scrutinized for obsolescence or changes in word meaning over a 13-year span. In each case, no apparent justification was found for the observed differential changes in difficulty. Since the plot correlations are no lower than .96 and relatively few items were found to evidence appreciable differential changes in difficulties, it seems likely that such differential changes have little effect, if any, on either the mean scores of the experimental samples or the equating results.

Equating Results

Raw scores on two pairs of 1963 and 1973 tests that were administered in counterbalanced order were equated by setting means and standard deviations equal after making the necessary adjustments for practice or fatigue effect. As will be noted from the summary statistics in Table 2, the conditions for the linear equating method of this study are satisfied: the four pairs of tests are parallel forms with minimal observed differences in their reliabilities, standard errors of measurement, difficulty indices and distributional characteristics. The raw-score distributions on the pair of tests taken by each sample are given in Appendixes VII to X.

In adjusting for practice or fatigue effect for the Design II equating method, a slight fatigue effect was observed for all four samples. This was estimated to result in a depressed raw-score of .64 to .68 of a point (or about 4 scaled-score points) on the test taken second for each of the samples.

The slopes and intercepts of the operational conversion lines used for placing the March and November 1973 SAT-verbal raw scores on the reporting scale as well as those of the experimentally derived conversion lines for transforming the raw scores on these tests to their equivalent scores on the May and December 1963 scales, respectively, are presented in Table 3. For a given reported score on the March or November 1973 test, the equivalent score on the corresponding May or December 1963 scale was computed from these conversion parameters. The two experimentally derived scores on the 1963 scales were then averaged to yield a one-to-one correspondence between a 1973 reported score and its equivalent score on the 1963 scale.

Presented in Table 4 are the scores at specific points on the 1973 reported-score scale and their equivalent linear and curvilinear experimentally derived scores on a 1963 scale for each of the 1973 tests, as well as for the average of the two 1973 tests. The linear equating results indicate that a performance level of 800 as reported in 1973 corresponds to an average of 803 on the rederived scale for the two 1973 forms; similarly, a reported score of 600 corresponds to 596 on the experimentally derived scale; a reported score of 450 corresponds to 441; and a reported score of 300 corresponds to 286.

Thus, at the reported score levels of 600 to 800 there was a maximum possible shift in scale of 4 points. However, as the reported score level decreases below 600, the scale shift increases to a maximum of 18 points for the average of the two forms at the 200 score level. These results are illustrated in the graph of the 1973 reported scores compared with that of their equivalent scores on the experimentally derived 1963 scale (see Appendix XI).

As presented in Table 5, the reported mean score of 453 for the four experimental samples corresponds to a derived mean score of 444 for a scale shift of 9 points over the 10-year period for the average of the two experimental equatings. That is, the mean reported score was 9 points higher, as determined from the study, than it would have been had the scale remained invariant. The actual candidate groups tested in March and November 1973, totaling 500,965 candidates, obtained an average reported mean score of 454 compared to 453 for

Table 2. Summary Statistics for Four Samples (A-D)
That Took Pairings of 1963 and 1973 SAT-Verbal in Counterbalanced Order[a]

Sample	N	SAT-verbal test[a]	Raw scores				r_{xy}[b]	Reliability (K-R#20)	Standard error of measurement	Difficulty index (delta)[c]		R-Biserial	
			Mean	S.D.	Range	Skewness				Mean	S.D.	Mean	S.D.
A . . .	790	May 1963 (Y1)	34.1	16.9	-10-86	0.15	.90	.92	4.9	13.1	2.6	0.46	0.11
		March 1973 (X1)	34.1	16.1	- 5-80	0.31	(.98)	.91	4.8	13.1	2.7	0.45	0.12
B . . .	790	March 1973 (X1)	35.7	17.1	- 6-83	0.28	.92	.92	4.8	12.9	2.7	0.47	0.12
		May 1963 (Y1)	34.5	17.6	- 2-83	0.29	(1.00)	.92	4.9	13.0	2.6	0.48	0.11
C . . .	785	December 1963 (Y2)	35.1	16.8	- 5-83	0.17	.90	.92	4.9	13.0	2.6	0.46	0.12
		November 1973 (X2)	34.9	16.1	- 5-80	0.24	(.98)	.91	4.8	13.0	2.8	0.45	0.13
D . . .	795	November 1973 (X2)	36.0	16.0	- 1-79	0.22	.90	.91	4.8	12.9	2.8	0.44	0.13
		December 1963 (Y2)	34.9	16.9	- 4-81	0.23	(.98)	.92	4.9	13.0	2.5	0.45	0.11

a. Presented in the order in which the tests were actually taken.
b. Correlation between scores on the 1963 and 1973 tests, with the correlation corrected for attenuation in parentheses.
c. The difficulty index, delta, is the percent correct for each item expressed as a normal deviate on a linear scale with a mean of 13.0 and a standard deviation of 4.0. The higher the delta value, the more difficult the item.

Table 3. Parameters for Placing 1973 SAT-Verbal Scores on the 1963 and 1973 Scales

	Conversion parameters				
	1973 scale (operational)		1963 scale (experimental)		Possible scaled-score range
Test form	Slope	Intercept	Slope	Intercept	
March 1973	6.3059	231.6546			93-799
			6.5404	218.7723	75-807
November 1973 . . .	6.3281	229.9625			91-799
			6.5219	209.5446	66-797

the experimental samples. One can, therefore, conclude that the experimental samples represent the same ability level as the actual 1973 candidate groups.

If the two pairs of equating results had not been averaged but had been considered separately, it can easily be observed from Table 4 that the results for the November 1973 test produced a larger scale shift than was evidenced from the March 1973 equating results. Thus, a discrepancy in the equating lines resulting from the choice of a specific pair of forms to be equated does exist. As indicated above in the design of the study, the ideal method of establishing the scale shift between 1963 and 1973 would have been to equate the scores on all possible pairings of the four 1963 and five 1973 forms and to have averaged the results for the 20 pairings of equatings that would have been administered to 40 samples in counterbalanced order.

Had the May 1963 and the March 1973 forms only been used for the experimental equatings, it would have been concluded (see Table 4) that an average score of about 453 on the 1973 scale had shifted upward from the 1963 scale by about 5 points. By comparison, the experimental equating of the December 1963 and November 1973 test forms resulted in an upward scale shift of about 14 points at the same 1973 reported-score level of 453.

The results of the curvilinear equatings, also presented in Table 4, closely parallel the linear results. Comparisons of the average of the linear with the average of the curvilinear equating results for the two 1973 forms reveal a discrepancy of at most 5 points at the extremities of the score scales (ignoring the extrapolated values in parentheses) but no difference at the 500 score level. The linear and the equipercentile results are consistent in the middle range of the score scale.

Table 4. Comparison of 1973 Reported Scores with Equivalent Scores on the 1963 Experimental Derived Scale

1973 reported score	Equivalent score on 1963 scale for Form X1 (March 1973 test)		Equivalent score on 1963 scale for Form X2 (November 1973 test)		Equivalent score on 1963 scale for average of two 1973 tests	
	Linear scale	Curvilinear scale	Linear scale	Curvilinear scale	Linear scale	Curvilinear scale
800	808	(785)[a]	797	(781)[a]	803	(783)[a]
750	756	(734)[a]	746	(732)[a]	751	(733)[a]
700	705	700	694	688	699	694
650	653	654	642	639	648	646
600	601	599	591	589	596	594
550	549	550	539	545	544	547
500	497	496	488	488	492	492
450	445	448	436	438	441	443
400	393	397	385	383	389	390
350	342	339	333	332	337	335
300	290	288	282	281	286	285
250	238	233	230	225	234	229
200	186	(194)[a]	179	(182)[a]	182	(188)[a]

a. Figures in parentheses were obtained by extrapolation from graphed data.

Table 5. Summary Statistics on 1973 Tests for Reported vs Experimentally Derived Score Scales for the Study Samples

Test form	Sample[a]	Size of sample	1973 reported score		Experimentally derived linear scale	
			Mean	Std. Dev.	Mean	Std. Dev.
March 1973	A	790	446.5	101.6	441.6	105.4
	B	790	456.9	107.8	452.4	111.8
November 1973 . .	C	785	450.6	101.7	436.9	104.8
	D	795	457.8	101.4	444.4	104.5
Average of two tests	A-D	3,160	453.0	103.1	443.8	107.0

a. Samples A and C took the 1963 test first and the 1973 test second; samples B and D took the 1973 test first and the 1963 test second.

Discrepancy Between the 1963 and the March and November 1973 Scales

It seems appropriate at this point to explore in greater detail the discrepancy between the average scale drifts of about 5 and 14 points reported for the March 1973 and November 1973 tests, respectively, by examining the results of the operational equating for the May and December 1963 tests to which the 1973 tests used in this study were experimentally equated. In form-to-form equating of operational tests, each new form of the SAT is equated to two old forms. The slopes and intercepts of the conversion lines produced from the two independent equatings are then averaged to yield one conversion line for linking the new form to the reporting scale for the old forms. Usually, the conversion lines from the two independent equatings should be quite close to each other in the middle of the score scale, thus providing an additional check on the consistency of the equating results. Comparisons of the May and December 1963 equating results are presented in Table 6.

At the reported-score level of 453 for the November 1973 test, the two conversion lines with which the December 1963 test was placed on the reporting scale produced equivalent scores of 440 and 439 on the scales for a 1960 and a 1962 form, respectively. However, at the same reported-score level of 453 for the March 1973 test, the two equating lines linking the May 1963 test to two earlier forms yielded equivalent scores of 454 and 443, respectively, on the scales for a 1960 and a 1962 form for a discrepancy of 11 points. The greater consistency between the results of the two independent equating lines for the December 1963 scale and the closeness of those results with the May 1963 (Line 2) equating rather than with the May 1963 (Line 1) equating suggest that the extent of the scale drift ought to be based

Table 6. Comparison of 1973 Reported Scores with Equivalent Scores on the May 1963 and the December 1963 Experimentally Derived Scales Based on Two Separate Equating Lines for Each 1963 Form

1973 reported score	Equivalent score on May 1963 scale for Form X1 (March 1973 test)			Equivalent score on December 1963 scale for Form X2 (November 1973 test)			
	May 1963 equating to 1960 form (Line 1)	May 1963 equating to 1962 form (Line 2)	May 1963 scale from bisector of Lines 1&2	Dec. 1963 equating to 1960 form (Line 1)	Dec. 1963 equating to 1962 form (Line 2)	Dec. 1963 scale from bisector of Lines 1&2	Equivalent score on 1963 scale[a] for average of two 1973 tests
800	824	793	808	785	810	797	803
750	771	743	756	735	757	746	751
700	717	692	705	686	703	694	699
650	664	642	653	636	649	642	648
600	611	592	601	586	596	591	596
550	557	541	549	537	542	539	544
500	504	491	497	487	489	488	492
450	451	440	445	437	435	436	441
400	397	390	393	388	382	385	389
350	344	339	342	338	328	333	337
300	291	289	290	288	275	282	286
250	237	238	238	239	221	230	234
200	184	188	186	189	168	179	182
Mean 453	454	443	448	440	439	439	444

a. Average of the two bisector lines from equivalent scores on May and December 1963 scales.

on the May 1963 (Line 2) and the December 1963 (Lines 1 and 2) equatings, which produced more consistent results. Consequently, the larger upward scale drift of about 14 points produced for the November 1973 scale with respect to the December 1963 equatings appears to be a more reliable estimate than the 5-point shift observed for the March 1973 scale at the reported mean-score level of 453 for the 1973 candidate group.

If the equating results for conversion Line 1 of the May 1963 test were ignored and the reporting of the May 1963 score based only on the second conversion line, the reported score of 453 on the 1973 scale would have been equivalent to an average score of 441 on the 1963 scale. Thus, the upward drift for the 1973 scale would have amounted to an average of 12 points instead of the 9-point difference between 453 and its equivalent score, on the 1963 scale, of 444 produced by using both pairs of 1963 operational equating lines.

Equating Results for High- and Low-Ability Groups

In order to examine whether the equating results based on the experimental samples would vary if their average ability levels had been different from those of actual SAT candidate groups, additional equatings were performed for two subsamples of each experimental group. Each of the four experimental samples was categorized into two subsamples on the basis of their self-reported class ranks as indicated in Table 7. It is observed that the mean score difference between the high- and low-ability groups is approximately one standard deviation of the scores for the total group for each test. Despite the substantial difference between the two ability levels, the conversion parameters derived from the separate equatings based on the performance of high- and low-ability groups produced results similar to those obtained for the total group.

The raw-score means and standard deviations for the total group in each of the four samples were then converted to the experimentally derived scales by applying the parameters obtained from the three separate equatings. The resulting converted scores and parameters are presented in Table 8. It is noted that within each sample, the total-group mean scores and standard deviations based on the different parameters do not vary by more than 2 points. It is, therefore, reasonable to conclude that our study results would not have been affected had the ability levels of the SAT candidate groups been higher or lower than those of the experimental groups (who, incidentally, were found not to have been markedly different from the combined group of actual SAT candidates who took the March and November 1973 SAT-verbal).

Equating Error

In interpreting the observed scale shift between the 1973 reported and the 1963 experimentally derived score scales, consideration must be given to the amount of sampling error in the equating operation. The sampling errors for equating (Lord, 1950) associated with selected scores on the 1973 reporting scale are presented in Table 9 for

Table 7. Summary Raw-Score Data on 1963 and 1973 Verbal Tests for Total vs. High- and Low-Ability Groups[a]

Sample[b]	Ability level	N	Mean	Std. Dev.	Mean	Std. Dev.
			May 1963 test		March 1973 test	
A	High	319	44.8245	14.6210	44.4138	14.0062
	Low	456	27.3355	13.9386	27.4671	13.3176
	Total group	790	34.1241	16.9279	34.0634	16.1157
B	High	335	44.7075	15.7873	45.7701	15.3877
	Low	444	26.9167	14.9392	28.3874	14.1779
	Total group	790	34.4848	17.6199	35.7266	17.0928
			December 1963 test		November 1973 test	
C	High	315	45.3778	14.4585	44.7524	13.9944
	Low	457	28.3895	14.3585	28.4573	13.4222
	Total group	785	35.0803	16.8038	34.8599	16.0744
D	High	319	45.7022	14.7611	45.8777	14.0219
	Low	467	27.7152	14.0154	29.4261	13.6023
	Total group	795	34.9358	16.8781	36.0101	16.0283

a. High are those students who reported their class rank to be in the top fifth. Low are those students who reported their class rank to be below the top fifth. Total group includes those students who did not indicate their class rank.
b. Samples A and C took the 1963 test first; samples B and D took the 1973 test first.

the linear equating of the March and November 1973 forms to the May and December 1963 forms, respectively. The sampling errors associated with the equivalent scores on the 1963 linear scale for the average of the two 1973 tests (see Table 4) are also presented in the last column of Table 9.

On the basis of these estimated errors, a reported score of 750 on the 1973 scale is equivalent to a scaled score ranging from about 748 to 753 (i.e., 751 $\pm$ 2.5) on the 1963 scale for a discrepancy of -2 to +3 points at that level between the 1973 and 1963 scales. This observed discrepancy represents hardly any shift at that score level. Similarly, a reported score of 600 on the 1973 scale is equivalent to a scaled score ranging from about 594 to 597 on the 1963 scale for a discrepancy of 3 to 6 points in favor of the candidate of this verbal ability level who took the 1973 test. Applying the estimated error of about one score point at the 500 score level on the 1973 reporting scale yields an equivalent score of 492 $\pm$ 0.9 on the 1963 scale for a scale shift of 7 to 9 points in favor of a 1973 candidate of the same ability. Similarly, a mean score of 453 on the 1973 reporting scale, which is equivalent to an experimentally derived score of 444 on the 1963 scale, yields a discrepancy of 8 to 10 points between the two

scales when the equating error of about 1 score point at this level is taken into account. At the 300 score level, the discrepancy or shift between the 1973 and 1963 scales ranges from about 13 to 16 points in favor of the 1973 candidate. Thus, the results in Table 4 indicate almost no scale drift at the 650 score level and above, but drifts ranging from 2 scaled-score points at the 600 verbal score ability level to almost 20 score points at the 200 score level in favor of the 1973 candidates compared to the scores that 1963 candidates of corresponding ability levels had earned on the 1963 SAT-verbal.

Table 8. Conversion Parameters Derived from Separate Equating Results for Different Ability Subgroups and the Corresponding Total-Group Statistics Based on Those Parameters

Sample	Group used for parameters	Experimentally derived conversion line[a]		Total-group statistics using different parameters		Possible scaled-score range
		Slope	Intercept	Mean	Std. Dev.	
A . . .	High ability	6.5029	219.5720	441	105	77-805
	Low ability	6.6052	217.4121	442	106	72-812
	Total group	6.5404	218.7723	442	105	75-807
B . . .	High ability	6.5029	219.5720	452	111	77-805
	Low ability	6.6052	217.4121	453	113	72-812
	Total group	6.5404	218.7723	452	112	75-807
C . . .	High ability	6.4834	212.3194	438	104	70-796
	Low ability	6.5268	208.5140	436	105	65-796
	Total group	6.5219	209.5446	437	105	66-797
D . . .	High ability	6.4834	212.3194	446	104	70-796
	Low ability	6.5268	208.5140	444	105	65-796
	Total group	6.5219	209.5446	444	105	66-797

a. Conversion parameters for corresponding ability groups in Samples A and B, and C and D are identical because Design II linear equating combines group statistics for counterbalanced pairings of tests.

Table 9. Standard Error[a] of Equated Score

1973 SAT-verbal reported score	Experimentally derived verbal score on 1963 scale		
	March 1973 test	November 1973 test	Average for March and November 1973 tests
750	756 ± 3.5	746 ± 3.6	751 ± 2.5
600	601 ± 2.0	591 ± 2.1	596 ± 1.5
500	497 ± 1.3	488 ± 1.3	492 ± 0.9
400	393 ± 1.3	385 ± 1.3	389 ± 0.9
300	290 ± 2.0	282 ± 2.1	286 ± 1.5
200	186 ± 3.0	179 ± 3.2	182 ± 2.2

a. The estimated error due to sampling variance.

CONCLUSION

The results of this study support the conclusion reached in the 1975 study by the present authors that the SAT-verbal scale has drifted upward over the 10-year period, 1963-1973, but the magnitude of the drift is less than was estimated in the previous study. The experimental samples obtained a mean score of 453 on the 1973 verbal test forms that was found to be about 9 points higher than the mean score of 444 they would have earned on the 1963 reporting scale. It was observed, however, that little or no scale shift occurred in the score range of 650 and above.

The 9-point scale shift was also found to be smaller than the 16-point upward shift obtained in the previous study at the 1973 reported-score level of 453, which was then determined (from the results of the common-item equating of the December 1963 test form to that of December 1973) to be equivalent to 437 on the 1963 scale. Had the two independent equatings of the present study not been averaged, the scale shift at the 1973 mean-score level of 453 would have amounted to an upward drift of only about 5 points for the equating of one pair and about 14 points for the equating of the other pair of tests. Thus, there is an obvious difference in the equating results associated with the particular pair of tests selected for form-to-form equating. If one of the two conversion lines from the May 1963 operational equating that appears to provide inconsistent equating results with the second conversion line were ignored and only the line that provides more consistent results with those of the December 1963 equatings were considered, the scale drift would be about 11 to 13 points (taking into account a sampling error of about 1 point) rather than 8 to 10.

The implication of this study for the SAT-verbal score decline is that operational equatings during the 1963-1973 period have had the

cumulative effect of making the 1973 candidate groups appear more able than the 1963 candidates of comparable ability. Thus, the study results suggest that the 1973 reported scores underestimate the extent of mean verbal score decline over the 10-year period by about 8 to 10 points, although there is evidence that the scale shift could be as high as 11 to 13 points.

REFERENCES

Angoff, W. H., "Scales, Norms and Equivalent Scores," in R. L. Thorndike (ed.) Educational Measurement (Second Edition). Washington, D.C.: American Council on Education, 1971

Lord, F. M., Notes on Comparable Scales for Test Scores. Research Bulletin 50-48. Princeton, N.J.: Educational Testing Service, 1950

Modu, C. C., and Stern, J., The Stability of the SAT Score Scale. College Entrance Examination Board Research and Development Report 74-75, No. 3. Princeton, N.J.: Educational Testing Service, 1975

Stewart, E. E., The Stability of the SAT-Verbal Score Scale. College Entrance Examination Board Research and Development Report 66-7, No. 3. Princeton, N.J.: Educational Testing Service, 1966.

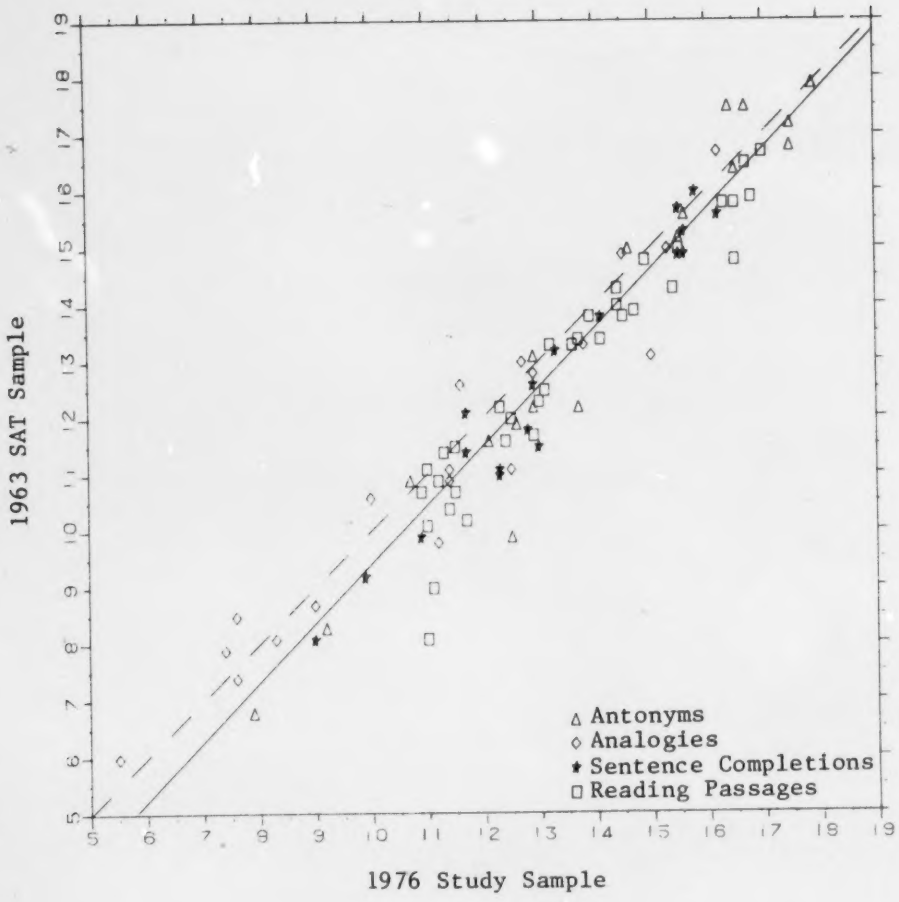
Appendix I. Description of 1963 and 1973 SAT-Verbal Sections on Which Scores Were Reported

Section	Item types	Number of items	
30-minute	Antonyms	10	40
	Analogies	10	
	Sentence Completion	10	
	Reading Passages	10	
45-minute	Antonyms	8	50
	Analogies	9	
	Sentence Completion	8	
	Reading Passages	25	
Total test			90

Note: All test forms were formula-scored (R-W/4).

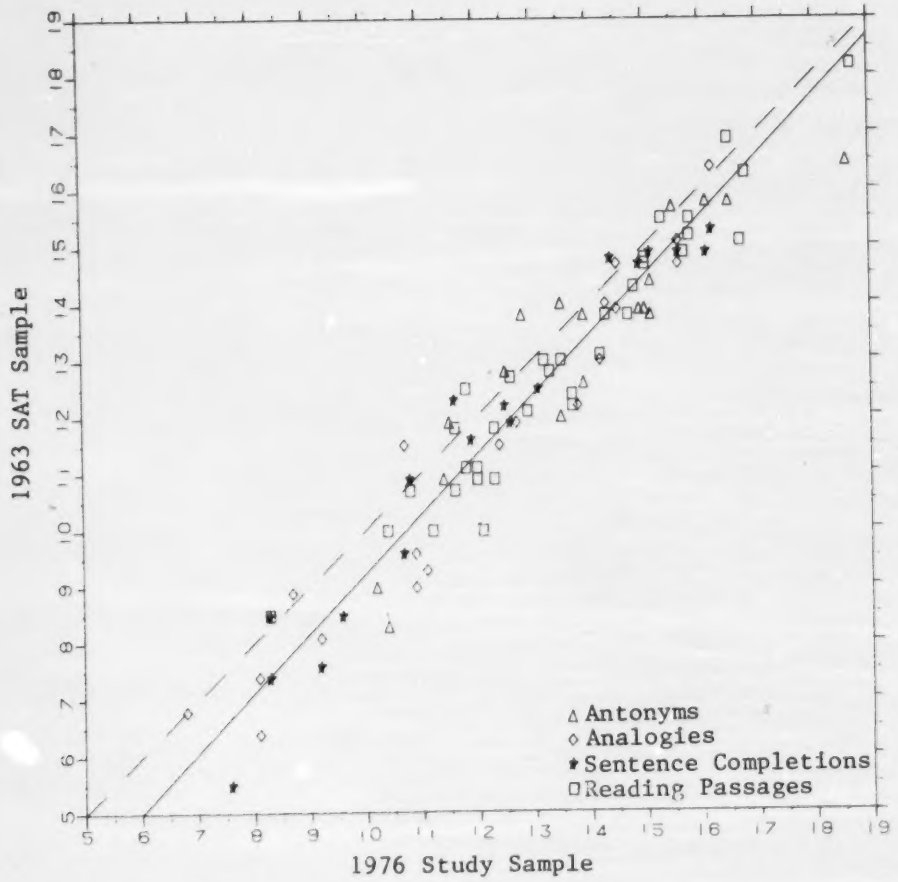
Appendix II. Delta Plot of May 1963 Verbal Items
for 1963 SAT Sample vs 1976 Study Sample

Sample A
R = 0.964


1963 SAT Sample
1976 Study Sample
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
△ Antonyms
◇ Analogies
★ Sentence Completions
□ Reading Passages

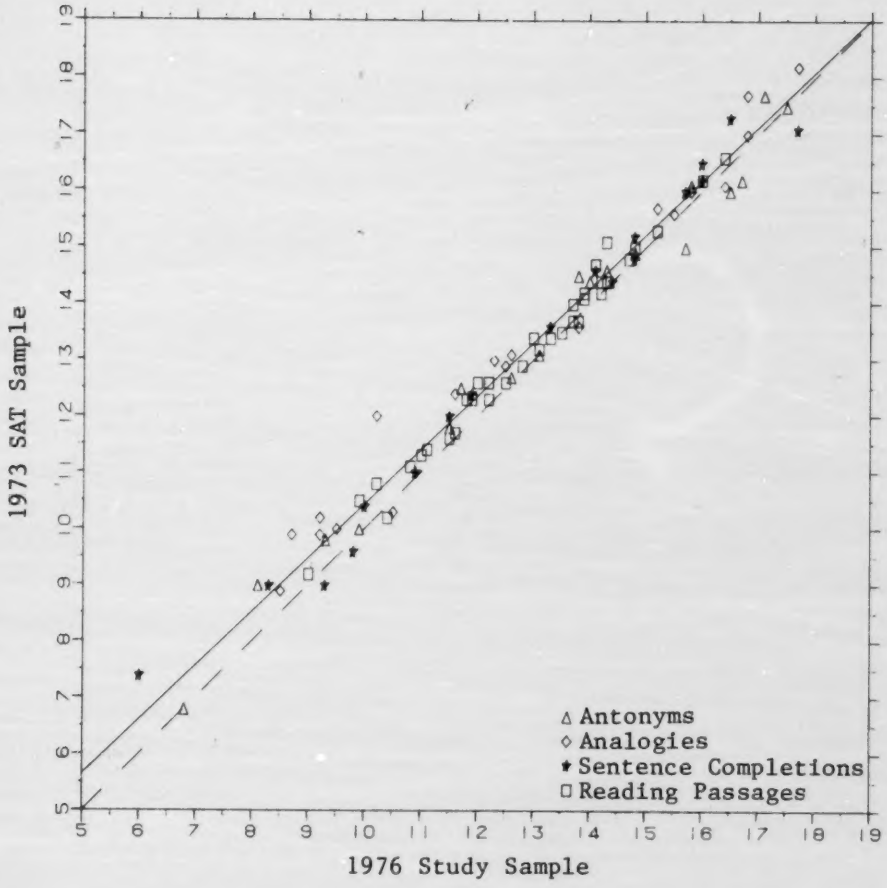
Appendix III. Delta Plot of December 1963 Verbal Items
for 1963 SAT Sample vs 1976 Study Sample

Sample C
R = 0.964


1963 SAT Sample
1976 Study Sample
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
△ Antonyms
◇ Analogies
★ Sentence Completions
□ Reading Passages

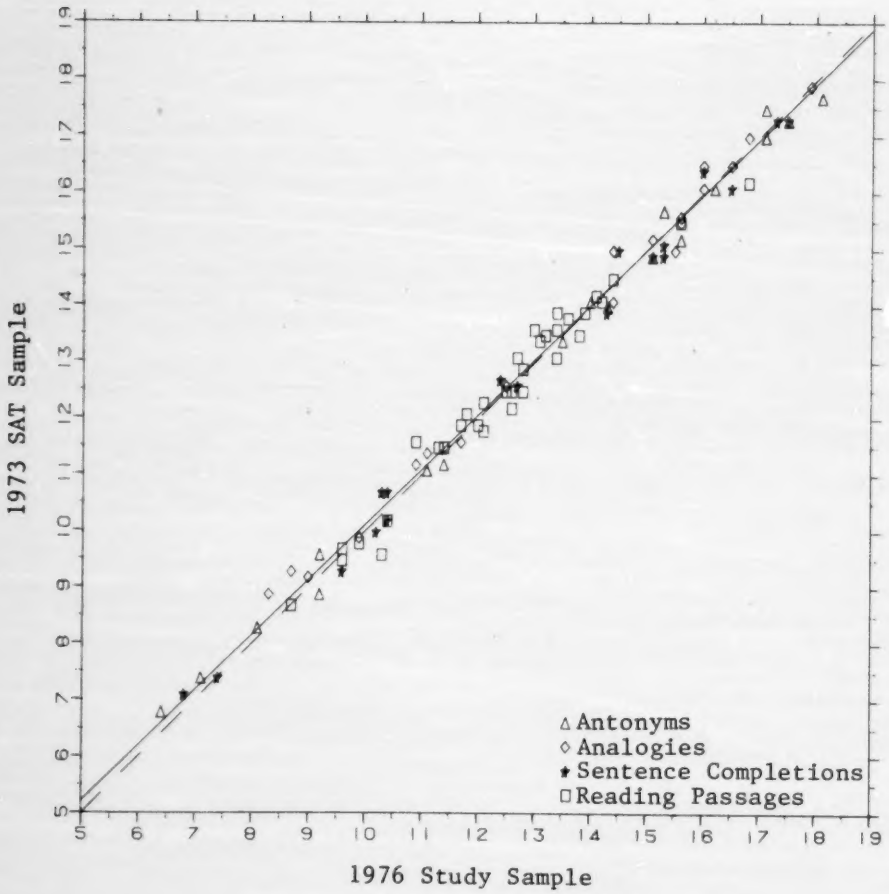
Appendix IV. Delta Plot of March 1973 Verbal Items
for 1973 SAT Sample vs 1976 Study Sample

Sample B
R = 0.989



Appendix V. Delta Plot of November 1973 Verbal Items
for 1973 SAT Sample vs 1976 Study Sample

Sample D
R = 0.994



Appendix VI. Question on Class Rank and the Frequencies of Students in Each Response Category

Question: How would you compare your achievement in subjects taken since beginning ninth grade with that of other students in your high school class?
(A) Top fifth
(B) Second fifth
(C) Third fifth
(D) Fourth fifth
(E) Lowest fifth

Frequencies of Obtained Responses

Class rank	Study samples			
	A	B	C	D
Top fifth	319	335	315	319
Second fifth	256	242	250	250
Middle fifth	167	170	175	175
Fourth fifth	27	28	26	36
Lowest fifth	6	4	6	6
No response	17	15	17	13
Total number	792	794	789	799

Appendix VII. Raw-Score Frequency Distributions of 1963 and 1973 Verbal Tests[a]

Sample A

May 1963 form			March 1973 form		
Raw score X	f	Percentile rank of lower limit of interval	Raw score X	f	Percentile rank of lower limit of interval
85. . . . 86	1	99.9	80. . . .	2	99.7
80. . . . 84	0	99.9	75. . . . 79	3	99.4
75. . . . 79	4	99.4	70. . . . 74	7	98.5
70. . . . 74	10	98.1	65. . . . 69	22	95.7
65. . . . 69	17	95.9	60. . . . 64	24	92.7
60. . . . 64	29	92.3	55. . . . 59	37	88.0
55. . . . 59	36	87.7	50. . . . 54	54	81.1
50. . . . 54	60	80.1	45. . . . 49	61	73.4
45. . . . 49	54	73.3	40. . . . 44	63	65.4
40. . . . 44	85	62.5	35. . . . 39	78	55.6
35. . . . 39	83	52.0	30. . . . 34	101	42.8
30. . . . 34	83	41.5	25. . . . 29	106	29.4
25. . . . 29	91	30.0	20. . . . 24	85	18.6
20. . . . 24	72	20.9	15. . . . 19	61	10.9
15. . . . 19	54	14.1	10. . . . 14	46	5.1
10. . . . 14	56	7.0	5. . . . 9	26	1.8
5. . . . 9	32	2.9	0. . . . 4	11	0.4
0. . . . 4	16	0.9	- 5. . . .- 1	3	0.0
- 5. . . .- 1	5	0.3		790	
-10. . . .- 6	2	0.0			
	790				

M_x = 34.12	M_x = 34.07
σ_x = 16.93	σ_x = 16.12
Md_x= 33.61	Md_x= 32.50
n = 90 items	n = 90 items

a. Ordered as taken by the sample.

Appendix VIII. Raw-Score Frequency Distributions of 1973 and 1963 Verbal Tests[a]

Sample B

March 1973 form			May 1963 form		
Raw score X	f	Percentile rank of lower limit of interval	Raw score X	f	Percentile rank of lower limit of interval
80. . . . 83	7	99.1	80. . . . 83	2	99.7
75. . . . 79	4	98.6	75. . . . 79	11	98.4
70. . . . 74	11	97.2	70. . . . 74	12	96.8
65. . . . 69	26	93.9	65. . . . 69	24	93.8
60. . . . 64	31	90.0	60. . . . 64	28	90.3
55. . . . 59	35	85.6	55. . . . 59	39	85.3
50. . . . 54	54	78.7	50. . . . 54	45	79.6
45. . . . 49	62	70.9	45. . . . 49	63	71.6
40. . . . 44	77	61.1	40. . . . 44	67	63.2
35. . . . 39	99	48.6	35. . . . 39	79	53.2
30. . . . 34	80	38.5	30. . . . 34	94	41.3
25. . . . 29	80	28.4	25. . . . 29	94	29.4
20. . . . 24	89	17.1	20. . . . 24	67	20.9
15. . . . 19	56	10.0	15. . . . 19	57	13.7
10. . . . 14	34	5.7	10. . . . 14	51	7.2
5. . . . 9	27	2.3	5. . . . 9	36	2.7
0. . . . 4	15	0.4	0. . . . 4	17	0.5
- 5. . . .- 1	2	0.1	- 2. . . .- 1	4	0.0
- 6. . . .	1	0.0			
	790			790	
M_x = 35.73			M_x = 34.48		
σ_x = 17.09			σ_x = 17.62		
Md_x= 34.98			Md_x= 32.95		
n = 90 items			n = 90 items		

a. Ordered as taken by the sample.

Appendix IX. Raw-Score Frequency Distributions of 1963 and 1973 Verbal Tests[a]

Sample C

December 1963 form			November 1973 form		
Raw score X	f	Percentile rank of lower limit of interval	Raw score X	f	Percentile rank of lower limit of interval
80. . . . 83	2	99.7	80. . . .	3	99.6
75. . . . 79	4	99.2	75. . . . 79	5	99.0
70. . . . 74	13	97.6	70. . . . 74	8	98.0
65. . . . 69	13	95.9	65. . . . 69	13	96.3
60. . . . 64	31	92.0	60. . . . 64	27	92.9
55. . . . 59	42	86.6	55. . . . 59	43	87.4
50. . . . 54	60	79.0	50. . . . 54	49	81.1
45. . . . 49	74	69.6	45. . . . 49	69	72.4
40. . . . 44	69	60.8	40. . . . 44	80	62.2
35. . . . 39	87	49.7	35. . . . 39	88	51.0
30. . . . 34	79	39.6	30. . . . 34	89	39.6
25. . . . 29	76	29.9	25. . . . 29	88	28.4
20. . . . 24	82	19.5	20. . . . 24	79	18.3
15. . . . 19	63	11.5	15. . . . 19	69	9.6
10. . . . 14	43	6.0	10. . . . 14	34	5.2
5. . . . 9	31	2.0	5. . . . 9	33	1.0
0. . . . 4	13	0.4	0. . . . 4	3	0.6
- 5. . . .- 1	3	0.0	- 5. . . .- 1	5	0.0
	785			785	
M_X = 35.08			M_X = 34.86		
σ_X = 16.80			σ_X = 16.07		
Md_X= 34.66			Md_X = 34.00		
n = 90 items			n = 90 items		

a. Ordered as taken by the sample.

Appendix X. Raw-Score Frequency Distributions of 1973 and 1963 Verbal Tests[a]

Sample D

November 1973 form			December 1963 form		
Raw score X	f	Percentile rank of lower limit of interval	Raw score X	f	Percentile rank of lower limit of interval
75. . . . 79	6	99.2	80. . . . 81	1	99.9
70. . . . 74	12	97.7	75. . . . 79	10	98.6
65. . . . 69	18	95.5	70. . . . 74	7	97.7
60. . . . 64	28	91.9	65. . . . 69	20	95.2
55. . . . 59	46	86.2	60. . . . 64	36	90.7
50. . . . 54	56	79.1	55. . . . 59	32	86.7
45. . . . 49	78	69.3	50. . . . 54	64	78.6
40. . . . 44	82	59.0	45. . . . 49	67	70.2
35. . . . 39	77	49.3	40. . . . 44	69	61.5
30. . . . 34	88	38.2	35. . . . 39	78	51.7
25. . . . 29	93	26.5	30. . . . 34	90	40.4
20. . . . 24	82	16.2	25. . . . 29	87	29.4
15. . . . 19	64	8.2	20. . . . 24	75	20.0
10. . . . 14	34	3.9	15. . . . 19	68	11.4
5. . . . 9	25	0.8	10. . . . 14	49	5.3
0. . . . 4	5	0.1	5. . . . 9	23	2.4
- 1. . . .	1	0.0	0. . . . 4	17	0.3
			- 4. . . .- 1	2	0.0
	795			795	
M_X = 36.01			M_X = 34.94		
σ_X = 16.03			σ_X = 16.88		
Md_X= 34.87			Md_X= 33.86		
n = 90 items			n = 90 items		

a. Ordered as taken by the sample.

Appendix XI. 1973 Reported Scores Compared to Their Equivalent Scores on the 1963 Scale